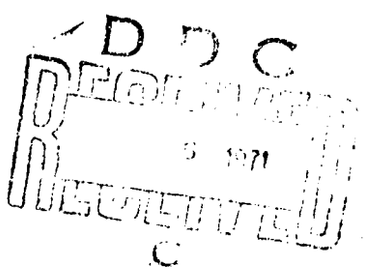
FTD-HT-23-473-70

FOREIGN TECHNOLOGY DIVISION

A SYNCHRONOUS SEARCH FOR DOCUMENTS

by

I. N. Kar-Yalayne



13

EDITED TRANSLATION

A SYNCHRONOUS SEARCH FOR DOCUMENTS

By: I. N. Kar-Yalayne

English pages: 9

Source: Seminar. Avtomatizatsiya Informatsionnykh Rabot i Voprosy Matematicheskoy Lingvistiki (Seminar on the Automation of Information Operations and Questions of Mathematical Linguistics), No. 2, 1966, pp. 66-75.

Translated by: H. Peck/NITHC

UR/0000-66-000-002

A SYNCHRONOUS SEARCH FOR DOCUMENTS

I. N. Kar-Yalayne
(Kiev)

In a complex document-data selective-expansion system the synchronous search unit is intended for solving special problems.

The preceding units of the system are designed for the preparation and input of source data to a computer, while the synchronous search unit distributes the abstracts of documents entering the data-search system according to subject-inquiries of interest to the customers (subscribers).

Each subject designation emerges as a bibliographical section for which the numbers of those documents whose contents correspond to the subject in the inquiry have to be compiled as a result of the synchronous search operation. In other words, there occurs a unique categorization of the incoming documents by a comparison of the search characteristics of the search claim (SC) with the search characteristics of the search pattern (SP).

To a certain extent, such a presentation of the problem predetermines its solution. The point is that in the selective data distribution system the problem of an incomplete output is more important than the problem of a superfluous output of documents corresponding to the subject presented. Therefore,

the operation of the search algorithm should be organized in such a manner as to remove data losses as much as possible.

In connection with this there arises the question of the criterion of the semantic coordination of our system. We consider that all documents having even the very minimum connection with it should be compiled for the subject-inquiry.

In other words, those documents are to be issued in which concepts are involved which have at least one element in common with the concepts named in the inquiry. We consider that SP corresponds to SC if at least one of the characteristics of SC is equal to one of the SP characteristics, i.e., if one of the vertices of the SC tree coincides with one of the vertices of the SP tree. Thus, the search claim is a disjunction of the search characteristics. On the one hand this assures minimum data losses (actually, due to some increase of the search noise), but on the other hand it maximally simplifies the search algorithm.

The type of semantic connection between the selected documents and the inquiry subjects succeeds in recognition without considerably complicating the algorithm in only one case – if the document and the inquiry are connected by an unconditional semantic generality [2]. Such a connection takes place if the convolution or first characteristic of the semantic code or SP scans of the document coincides with the SC convolution.

The unconditional semantic generality is recognized by a special unit. The documents selected by this unit are furnished with a "strong-operator" label and are singled out into a special list.

The realization of the selective-data distribution system requires the presence of the subjects according to which the documents are supposed to be distributed.

Unlike the inquiries which permit a reference inquiry into the system, the inquiries in the selective-data distribution system do not contain additional indicators and are expressed by one nominative word combination of the Russian language. For example, "Memories made of thin ferromagnetic films," "The investigation of the properties of superconductive films for the construction of digital machines," etc.

The list of the subjects by which the search is supposed to take place is considered as given. It is formed based on the requests coming from the system subscribers. However, the requested subject-inquiries have to be specially processed in a number of cases. In practice it often happens that some of the subjects which are of interest to various consumers are close to each other in a sense. Let us illustrate the above by the following example. Let us assume that there is supposed to be an inquiry on the following subjects:

1. The use of transistors.
2. The properties of transistors.
3. The technique of manufacturing transistors.

Let us now represent these subject names in the characteristics of the data language of the semantic codes. As a result we get three trees (see Fig. 1).

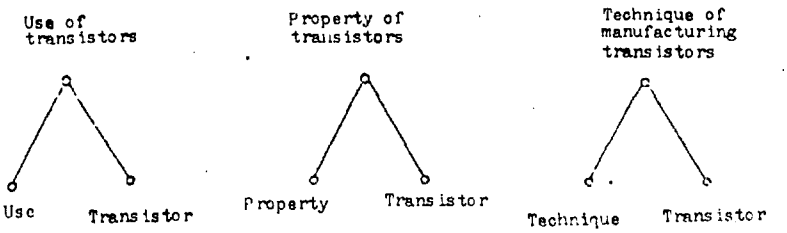


Fig. 1.

In the accepted method of searching for the disjunction of the search characteristics in all three cases the operation of the algorithm is reduced to selecting the documents with the

characteristics of "Transistor" (or in the search for such characteristics as "Use," "Properties," "Technique," there will appear extremely high search noise). Consequently, the same documents will be selected from these three subjects.

Thus, it is more expedient to combine similar subjects into one ("Transistor") to avoid the coincidence or excessively large intersection of quantities of documents corresponding to these subjects.

Let us examine some more such inquiry subjects such as "Magnetron," "Klystron," "Generator and the generation of superhigh frequency oscillations." In our operational dictionary we find the following definitions of the concepts encountered here. The magnetron is a "device which uses the properties of a constant magnetic field for creating the necessary electron trajectories. It is made up of a vacuum tube and a cavity resonator intended for generating superhigh-frequency oscillations" [2]. The klystron is an "electronic device made up of a tube and cavity resonator intended for amplifying and generating superhigh frequencies" [2]. The superhigh-frequency generator is a "device intended for generating superhigh-frequency oscillations." Even in the very surface analysis of these definitions it is absolutely obvious that these concepts are made similar by an important and essential characteristic – "the generation of superhigh-frequency oscillations." This fact clashes with the idea that obviously the subjects named can be combined into one – "Devices intended for generating superhigh-frequency frequencies." Obviously, combining the subjects in this manner is expedient to a certain definite limit. It is possible to present the following criterion which defines the degree of generalization.

Let the subscriber order subjects $T_1, T_2, \ldots, T_n$. We combine some of them into generalized subjects $T'_1 = T_{i_1} + T_{j_1} + \ldots + T_{k_1}$, $T'_2 = T_{i_2} + \ldots + T_{k_2}$, etc., those during whose search the search noise will not exceed the set value (let us

permit 50%) with respect to each of the requested subjects making them up. In other words, during the search of subject T_1' ("Devices intended for generating superhigh frequencies") no less than half the documents corresponding to subjects T_{i_1} ("Magnetron") and T_{j_1} ("Klystron") of each individually should be among the selected documents.

If it becomes impossible to reach the assigned search-noise value, even in a direct search for the requested subjects (such a position, for example, might be created in the search for the subject, "The technique of manufacturing transistors," since a large part of the selected documents will obviously be devoted to the use of transistors in circuits), it is expedient to make the search for the conjunction of the characteristics, i.e., to select a document with the simultaneous coincidence of two or several SC and SP characteristics (for example, "Transistor" and "Manufacture," or "Transistor" and "Production," etc.).

The system has been developed with the calculation for a multiaspect approach toward the search of documents. This means that the documents will be described from more than one viewpoint. Therefore, the necessary documents will be sought more successfully if the most essential search characteristics are focused in the inquiry.

In a synchronous search the search claim (SC) is compiled manually. The manual compilation formation of SC is inconvenient during a reference search, since it requires a large time expenditure, but it is advantageous in a selective data distribution: first, SC is formed for a prolonged period in this case; second, the manual formation gives the possibilities of verifying the quality of the search with various concepts of the same inquiry to establish experimentally what search characteristics to include in SC.

For example, let us show how the selection for the "Triode" SC search characteristics was done. First, such search characteristics as "vacuum tube," "cathode," "anode," "grid," and "vacuum device," i.e., characteristics corresponding to the special terms forming the meaning code were included in the composition of SC.

As a result of the search, together with such documents as "An analysis of the static regimes of amplifier triodes," "The restoration of the electric stability of thyratrons," "The investigation of ways to create nonfilament arc-discharge thyratrons," "Pulse thyratrons," etc., such documents were also selected which did not directly relate to the subject shown in the inquiry, for example: "A gas-discharge device with a hydrogen charge," "A reflex-klystron generator," "Magnetron harmonics on millimetric waves," etc. In this case the search noise appeared as a result of the fact that SC contained the wrong necessary search characteristics. Some of them had to be removed, first, the too general characteristic "vacuum tube," to limit the number of irrelevent documents as much as possible. From the experiments we performed we obtained the following results. A search for the subject "Generating superhigh-frequency oscillations" occurred from the characteristics of "generation" or "superhigh-frequency oscillations." With a synchronous search the following documents on this subject were selected.

1. "Magnetron harmonics on millimetric waves."
2. "Fixed-frequency generator."
3. "Millimetric-wave generators."
4. "The generation of millimetric oscillations by means of the Cerenkov radiation."
5. "Reflex-klystron generator for the 8-9 mm range."
6. "High-power magnetrons for the millimetric-wave range."
7. "Triode with a high characteristic curve used as a high-frequency amplifier, an oscillator and a frequency-mixer tube."

Since documents remotely associated in concept with the subject of the inquiry are taken into consideration during the synchronous search, we do not escape the search noise.

In our example it would be unavoidable if there were available documents associated with superhigh-frequency modulation and with the conversion of superhigh frequencies, etc., among the documents to be processed.

But avoiding the search noise after the removal of the "superhigh frequency" characteristic is impossible, because it would involve an inevitable loss of documents. At that time, an article under the title of "Superhigh-frequency tubes" would not be separated out since it can contain the necessary information. Other examples.

Subject: "High-frequency semiconductor devices."

Answer:

1. Transistor receiver for the 10-m range.
2. A 1 kW high-frequency transmitter.
3. Loss compensation in an oscillator circuit by means of a semiconductor triode.
4. New powerful ceramic triodes and tetrodes intended for very high frequencies.
5. New type of tungsten cathode for high-power oscillator tubes.
6. Semiconductor-triode transmitters.
7. Semiconductor-circuit triodes with low noise level for portable equipment.

Subject: "Magnetic recording by means of an electron beam."

The documents selected on this subject:

1. Magnetic recording by means of an electron beam.
2. Method of reducing the grid temperature in tubes.
3. The SRG-40 k receiver.
4. High-frequency diode with negative resistance.
5. A simple method of tuning IF circuits in the serial production of receivers.

6. The use of a magnetic recording.
7. Magnetic recording and reproduction.
8. The magnetic recording of a pulse amplitude.
9. Signal loss in a magnetic-tape recording.

Subject: "Methods of magnetic recording."
Selected documents:
1. The use of a magnetic recording.
2. Magnetic recording by an electron beam.
3. Magnetic recording and reproduction.
4. The magnetic recording of a pulse amplitude.
5. Signal loss in a magnetic-tape recording.

The algorithm has been realized on the "Minsk-2" computer. The program includes 400 octal commands and occupies approximately 550 octal cores.

The program consists of three units:

1. A self-search program.
2. A program for establishing the unconditional semantic generality between the selected documents and the subject-inquiries.
3. A program of the output of the results.

The input data for the program is an array of the scans of the search patterns and an array of the search claims. The output data are SC numbers for which the numbers of the sequential documents were compiled.

The search claims are stored on magnetic tape. Their number, i.e., the number of subjects for which synchronous searches are made, can reach several thousand, whereas the search for 500-600 subjects is made without additionally resorting to magnetic tape.

References

1. Gryaznukhina, T. A., E. F. Skorokhod'ko, and L. E. Pshenichnaya. Sistema informatsionnogo poiska (mashinnyy poisk

literatury) (Data search system (a reference machine search)), Izd-vo Naukova dumka, Kiev, 1964.

2. Khaykin, S. E. Slovar' radiolyubitelya (The radio amateur's dictionary), Gosenergoizdat, Moscow, 1960.

Received 25 December 1965

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Foreign Technology Division
Air Force Systems Command
U. S. Air Force

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3. REPORT TITLE

A SYNCHRONOUS SEARCH FOR DOCUMENTS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Translation

5. AUTHOR(S) (First name, middle initial, last name)

Kar-Yalayne, I. N.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
1969	9	2

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO. 6050205	FTD-HT-23-473-70
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d. DIA Task No. T68-05-02	

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Foreign Technology Division Wright-Patterson AFB, Ohio

13. ABSTRACT

An algorithm is described of a synchronous search in a complex system of selective retrieval of documents, with an allowance for exclusion of information loss. All documents containing if only one concept common with the concept specified in the claim slip are retrieved. The search pattern (SP) corresponds to the search claim (SC) if even one SC-characteristic is equal to one of the SP-characteristics. Then, SC is a disjunction of search characteristic, which greatly simplifies the search algorithm. Examples are given that illustrate the disjunction method of search. The algorithm was realized on a Minsk-2 digital computer. The program includes 400 octonary instructions and occupies 550 octonary cells. The program consists of three blocks: the search proper, the establishment of absolute community of meaning between selected documents and claims, and the delivery of results. An SP-file and an SC-bank serve as input information. The numbers of SC which correspond to the numbers of coming documents serve as output information; SC are stored on a magnetic tape. The number of topics served by the synchronous search goes into thousands; a search within 500-600 topics is performed without additional access to the tape. Orig. art. has: 1 figure. [AR8028559]

DD FORM 1 NOV 68 1473

UNCLASSIFIED
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Data Retrieval Automatic Document Analysis						